Download Genomic Sequences From NCBI

ftp://ftp.ncbi.nlm.nih.gov/genomes/

Locating and downloading assembly-based RefSeq genomic sequences for a set of organisms from NCBI FTP site

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Overview

Advances in technology have made large numbers of assembled genomes available. Submitted assemblies selected for NCBI Reference Sequence project (RefSeq) are annotated by NCBI's genome annotation pipeline and made available through the NCBI FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/). The hierarchical organization of the FTP directories, by taxonomic groups and then by species and assemblies, facilitates finding and downloading data for individual genome assemblies, but is not well suited for downloading data for all assemblies from a broad taxonomic group. Here, we describe a Linux shell command-based workflow that takes advantage of the assembly_summary.txt file for representative taxonomic groups, to extract URLs for files or directory of interest, and use them to download selected sequences or all data files.

Genomes FTP Directory Structure

The ***/genomes/refseq*** and ***/genomes/genbank*** directories organize available data by large taxonomic groups, i.e., archaea, bacteria, fungi, invertebrate, plant, protozoa, vertebrate_mammalian, vertebrate_other, and viral (last is RefSeq only, details at: http://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq/). Each group level directory contains an assembly_summary.txt file with details on the latest versions of assemblies available for that group. This file also contains many fields of metadata, useful in identifying genome assemblies of interest, as well as the URLs for the subdirectories from which the data files can be downloaded. For a detailed description of the file structure, see ftp://ftp.ncbi.nlm.nih.gov/genomes/README_assembly_summary.txt.

NCBI organizes the genomes data files with a consistent directory hierarchy. For example, the RefSeq entries have the following naming convention (GenBank entries have GCA instead of GCF initial): /all/GCF/aaa/bbb/ccc/GCF_aaabbbccc.V_NAME/GCF_aaabbbccc.V_NAME_X_Y.gz, where GCF_aaabbbccc.V is the assembly's accession.version, NAME is the assembly name, and _X_Y are sequence and file type. Workflows below use Linux shell utilities to process the assembly_summary.txt for a selected taxonomic group into FTP URLs for genomic sequences or full subdirectory content download.

Use Cases

Case 1: Get all the genomic sequence files for the fungal RefSeq assemblies

Under the ***/genomes/refseq*** directory of the NCBI FTP site, available data are grouped by large taxonomic groups, i.e., archaea, bacteria, fungi, invertebrate, plant, protozoa, vertebrate_mammalian, vertebrate_other, and viral, each with its own assembly_summary.txt file that provides detailed information of available assemblies along with the URLs for those subdirectories in the 20th column. The workflow consists of two steps, collecting and modifying the FTP URLs for the desired file format (genomic FASTA sequences), and downloading the relevant files using the collected URLs as input.

Step 1. Collect and modify the FTP URLs to point to the _genomic.fna.gz files

The command line is a pipe symbol– linked set:

```
curl 'ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/fungi/assembly_summary.txt' | \
awk '{FS="\t"}  !/^#/ {print $20} ' | \
sed -r 's|(ftp://ftp.ncbi.nlm.nih.gov/genomes/all/.+/)(GCF_.+)|\1\2/\2_genomic.fna.gz|' > genomic_file
```

The "\" is a Linux shell command to indicate that the command line continues in the next line. We use it to break the linked commands into distinctive steps so we can clearly see and discuss each sub-step:

- The first **curl** command simply gets the specified assembly_summary.txt file and passes its content to the next step with a pipe ("|", instead of displaying it in console).
- The second command uses the **awk** utility to separate each line's content by tab (FS="\t"), skip header line (!/^#/) and print out the value of the 20th column (print $20), and passes the output to the next step with pipe ("|").

NOTE: if you need the assembly submitted to GenBank, you will need to change the curl command's "refseq" to "genbank" and change sed command's "GCF" to "GCA" since their accession initials are different.

Use Cases (cont.)

- The third command uses **sed** to modify the extracted URL string that points to an assembly directory to point to the "_genomic.fna.gz" file instead. Specifically, with the pipe ("|") as delimiter, it first matches the URL into substrings using regular expression matching and captures them using parentheses (**s|(ftp://ftp.ncbi.nlm.nih.gov/genomes/all/.+/)(GCF_.+)|**), then reconstructs the string (**\1\2**) for the path and add another directory level (**/**) and a specific file name (**\2_genomic.fna.gz**). This modifies the existing URL to point to the **_genomic.fna.gz** file for that assembly. The last part (**>genomic_file**) redirects the output into a file named genomic_file (partially shown below).

```
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/002/945/GCF_000002945.1_ASM294v2/GCF_000002945.1_ASM294v2_genomic.fna.gz
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/149/845/GCF_000149845.2_SJ5/GCF_000149845.2_SJ5_genomic.fna.gz
```

Step 2. Use the output from step 1 as input to wget to download

The second step is very simple. The command (below left) calls the wget utility and passes the output from step 1 as an argument to the "—input-file" switch. It will iterate through the FTP URLs and pull down those files from NCBI's FTP site to the working directory (partially listed, below right).

```
wget --input genomic_file
```

```
-rw-r--r-- 1 samd sdesk  3989616 Dec 31 10:52 GCF_000002945.1_ASM294v2_genomic.fna.gz
-rw-r--r-- 1 samd sdesk  3492573 Dec 31 10:52 GCF_000149845.2_SJ5_genomic.fna.gz
```

The command gunzip *.gz will unpack them all to regenerate the FASTA files for further downstream need. For better file management, first move *.gz files to a new directory so they are isolated from other files.

Case 2: Get the directories and their contents for all the fungal RefSeq assemblies

If we wish to completely mirror and archive the files for this group or organisms, we can modify the above commands to download the directories along with all their contents (excluding subdirectories).

Step 1. Collect and modify the FTP URLs to get only the directory name

```
curl 'ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/fungi/assembly_summary.txt' | \
awk '{FS="\t"} !/^#/ {print $20"/"}'  > genomic_directory
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/149/845/GCF_000149845.2_SJ5/
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/150/505/GCF_000150505.1_SO6/
```

We can modify the command from Case 1 by dropping the sed command, and modify the awk slightly to get the directory URLs (left). Last two lines are example output.

Step 2. Pass the filte to wget to pull down the directories and their contents

```
wget -r --no-parent --no-host-directories --cut-dirs=2 --level=1 \
--input-file=genomic_directory
-r: recursively works through the directory
--no-parent: ignores the parent directory
--no-host-directories: saves the files without prepending the NCBI FTP URL
--cut-dirs=2: saves the files without creating intermediate directories
--level=1: works only at that level of directory
--input-file=value: sets directory input to the file specified by value
```

We use the wget command (left) to get all the directories and their files. The command is more complex than in Case 1, so we explain the meaning of each command line arguments separately below the command. This will pull all the directories and their contents down to your Linux box, so make sure you have enough disk space to handle them.

NOTE: For users on PCs without Linux or cygwin access, we can do the same first steps described above in a different way. The example at the end uses inline Perl commands (below). It allows us to generate the same outputs as on Linux, and then use the same wget commands for the second steps to download. The PC port of wget is available from: https://eternallybored.org/misc/wget/

```
perl -e "use LWP::Simple; $file=get(\"https://ftp.ncbi.nlm.nih.gov/genomes/refseq/fungi/assembly_summary.txt\"); while ($file
=~ /(ftp.+)(GCF_.+?)\s/g){print $1, $2, \"\/\", $2, \"_genomic.fna.gz\n\";}"  > fungi_genomic_files
perl -e "use LWP::Simple; $file=get(\"https://ftp.ncbi.nlm.nih.gov/genomes/refseq/fungi/assembly_summary.txt\"); while ($file
=~ /(ftp:.+GCF.+?)\s/g){print $1, \"\/\n\";}"  > fungi_directory
```

Warning: Type all example commands instead of copying/pasting since hidden characters may break their execution.